Spatial Sampling Design for Prediction with Estimated Parameters

Zhengyuan Zhu and Michael L. Stein[1]

[1]Zhengyuan Zhu is Assistant Professor, Department of Statistics and Operations Research, the University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (E-mail: zhuz@email.unc.edu). Michael L. Stein is Professor, Department of Statistics, the University of Chicago, Chicago, IL 60637. This research was supported in part by National Science Foundation grant DMS 99-71127.

Abstract

We study spatial sampling design for prediction of stationary isotropic Gaussian processes with estimated parameters of the covariance function. The key issue is how to incorporate the parameter uncertainty into design criteria to correctly represent the uncertainty in prediction. Several possible design criteria are discussed that incorporate the parameter uncertainty. A simulated annealing algorithm is employed to search for the optimal design of small sample size and a two-step algorithm is proposed for moderately large sample sizes. Simulation results are presented for the Matérn class of covariance functions. An example of redesigning the air monitoring network in EPA Region 5 for monitoring sulfur dioxide is given to illustrate the possible differences our proposed design criterion can make in practice.

KEY WORDS: Fisher information matrix, geostatistics, kriging, Kullback divergence, optimization, simulated annealing

1 Introduction

A common problem in spatial statistics is to observe a random process Z at a set of sample locations $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset D$, and then make inference about the unobserved $Z(x)$ for $x \in D$, where D is the region of interest. The network design problem of choosing the sample locations $S \subset D$ so that one can have the most accurate prediction (point prediction and/or prediction interval) in D is of great importance in many applications such as soil science, agriculture, and air pollution monitoring. Here we will use the redesigning of a Sulfur Dioxide (SO_2) monitoring network in the four mid-west states (IL, IN, OH, and KY) as a motivating example.

SO_2 is produced during the burning of sulfur-containing fuels such as coal and oil, during metal smelting, and by other industrial processes. It can affect the respiratory system, the functions of the lungs and irritate our eyes, causing coughing, mucus secretion, aggravating conditions such as asthma and chronic bronchitis and making people more prone to respiratory tract infections. The highest concentrations of sulfur dioxide are generally found near large fuel combustion sources. The EPA air quality standards for the annual mean of SO_2 is 0.03 parts per million (ppm). The standard also requires a maximum 24 hour mean less than 0.14ppm and maximum three hour mean less than 0.5ppm. In our study we will only consider the annual mean.

EPA has a network for monitoring the level of SO_2 in the aforementioned four states. The data from this network can be used to find regions that do not comply with the EPA standards. It can also be used in epidemiology studies to estimate the effects of SO_2 on human health. In the latter case it is of particular importance to give accurate predictions of the SO_2 level at locations with no monitoring stations, as well as accurate estimates of uncertainties in those predictions, in order to correctly assess the health effects of SO_2. The original network was designed to find the nonattainment area; thus, there is a large concentration of monitoring stations in areas where the

level of SO_2 is expected to be high, such as along the Ohio river valley. In recent years, however, there has been such a large reduction of SO_2 levels that almost no county in this region is in nonattainment, and it is of interest to reduce the monitoring network for SO_2 while maintaining its ability for accurate spatial prediction. We show in this paper how design criteria can be modified to achieve good spatial prediction when the covariance parameters have to be estimated from the same data, and develop efficient algorithms to find designs that are approximately optimal for such design criteria. The designs we get are different from the usual space filling designs in that there are small fractions of points that are closely located. Simulation studies show that such designs tend to give more accurate estimates of prediction error variance, while the variance of the point prediction error is similar to the traditional space filling designs.

A number of authors have investigated the problem of spatial sampling design assuming the parameters of the correlation function known, see for example McBratney et al. (1981) and Yfantis et al. (1987), who provide empirical evidence that when using kriging as the prediction method and the average or maximum kriging variance as the criterion, the equilateral triangular grid is apparently nearly optimal. Sacks et al. (1989) consider optimal sampling design for prediction in the context of computer experiments. Their model assumes no measurement error and the region of interest D is usually in a high dimensional space. Benhenni and Cambanis (1992) and Ritter (1996) consider designs for predicting the weighted integral of a stochastic process in $\mathbb{R}$ and show that sampling at the quantiles of a particular density yields asymptotically optimal predictors. Stein (1995) considers the problem of estimating the weighted integral of an isotropic random field $Z(x)$ over D in more than one dimension based on locally lattice sampling designs. He finds the asymptotically optimal cubature rules within the class of locally lattice designs and conjectures that these rules, which are based on designs that are locally an equilateral triangular lattice, are

asymptotically optimal with respect to all cubature rules based on point evaluations of Z.

In many applications, we have to use the same data for both estimation and prediction. Caselton and Zidek (1984), Caselton et al. (1992), Guttorp et al. (1993), and Zidek et al. (2000) developed the maximum entropy approach for designing monitoring networks by modeling observations at different monitoring locations as multivariate time series. Caselton et al. (1992) use a prior on the covariance matrix to incorporate the model uncertainty in the design criterion. However, since the processes are modeled as a multivariate time series, it is not clear how such designs perform for prediction at new locations. Wikle and Royle (1999) consider dynamic designs for space-time models, assuming that one can change the design at each new time. They estimate prediction error variances at time $t+1$ using observations up to time t, and use it to construct designs at time $t+1$. Banjevic and Switzer (2002) consider designs for processes with unknown variance function but known correlation function, and use a Bayesian approach to incorporate the uncertainty in estimating the parameters of the variance function. Most recently, Wiens (2004) considers designs that are robust against misspecified variance/covariance structures.

How to find the optimal static design for prediction with estimated covariance parameters remains a largely unexplored topic and will be the focus of this paper. Zimmerman (2005) considers designs that minimize the average approximate mean square errors of the empirical best linear unbiased predictors, which account for covariance parameter estimation uncertainty in the point prediction, but the impact on estimating the variance of prediction error is not addressed. We derive appropriate design criteria that account for estimation uncertainty in both point and interval prediction using asymptotic approximations. Simulation studies on small sample sizes demonstrate the usefulness of these criteria. For moderate to large sample sizes, the numerical difficulty in searching for the optimal spatial sampling design using any reasonable design criterion is over-

whelming. Ko et al. (1995) and Lee (1998) discussed an exact algorithm for maximum entropy sampling based on branch-and-bound methods that can find the exact solution that maximizes the entropy, and successfully solved a problem of selecting 13 from 27 sites. However, it is difficult to apply their approach to larger problems even if we could find a good bound for our criterion function. Thus we seek an approximate algorithm that can find a good local minimum and call it the best design instead of the optimal design. Simulated annealing (SA) is an optimization method that has been successfully applied to similar problems by van Groenigen and Stein (1998), Lark (2002), and Wiens (2004). However, the criterion we use for prediction with estimated parameters is more computationally demanding and depends on the spatial locations of the sampling points in a more complicated way. Although we have successfully implemented the SA algorithm for 30 observation designs, for sample sizes more than 100 SA cannot yield satisfactory results in reasonable time. Motivated by this numerical challenge we developed the two-step algorithm, which can be used for designs with moderately large sample sizes. It uses some of the sites to find the best design for prediction with known covariance parameters and then, conditional on those sites, uses the rest to find the best design for estimation of those covariance parameters. The proportion of the sites assigned for the two components is then varied to find the ratio that is best for prediction with estimated parameters. Using this method, the computationally more intensive criterion for prediction with estimated parameters only needs to be evaluated for the one dimensional optimization of the ratio, which dramatically simplifies the computation.

It is worth pointing out that the best designs found by the two-step algorithm are by no means unique. By changing the starting value in the first step, we can get designs that have similar criterion values but quite different spatial configurations, as long as the proportion of the sites for parameter estimation is kept constant. One could use the two-step algorithm to find several

designs for a given proportion, and then select a specific design based on other considerations, since in practice, most designs have to address multiple objectives, not just spatial prediction.

Section 2 presents the theory and methodology, including the spatial model, the general estimation and prediction methodology and the design criteria that will be used in later sections. Section 3 gives the two-step algorithm for finding approximately optimal designs, the numerical methods for simulation, and the methods used to compare designs using simulation. Section 4 contains the simulation results, Section 5 gives the air monitoring network example, and Section 6 discusses possible further work.

2 Theory and Methodology

We first give the spatial model assumed for the data, then briefly introduce kriging, the spatial prediction method, and Restricted Maximum Likelihood (REML), the parameter estimation method, together with the plug-in method used in practice for spatial prediction, and its asymptotic mean square prediction error (MSPE). At the end we introduce our design criterion for spatial prediction with estimated parameters.

2.1 Spatial Model

We focus on spatial sampling design for a stationary isotropic Gaussian model specified as follows. For $\mathbf{s}_i \in D$ and $i = 1, \ldots, n$, let $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$ and assume

$$\mathbf{Z} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{1}$$

where $\mathbf{X}$ is a known matrix of regressors of dimension $n \times p$, $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression coefficients, the i,jth element of $\boldsymbol{\Sigma}$ is $\sigma_{i,j}(\boldsymbol{\theta}) = \sigma^2\rho(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\xi}) + \tau^2\delta\{\mathbf{s}_i = \mathbf{s}_j\}$, δ being an indicator function with value 1 if $\mathbf{s}_i = \mathbf{s}_j$ and 0 otherwise, and $\boldsymbol{\theta} = (\boldsymbol{\xi}, \sigma^2, \tau^2)$. $\rho(\mathbf{u}; \boldsymbol{\xi})$ is

the correlation function with unknown parameter $\boldsymbol{\xi}$, and τ^2 is called the nugget variance in the geostatistics literature.

To carry out all the computations, it is convenient to work with a specific parametric family of correlation functions that is flexible yet simple. The Matérn family fits this description in that it has only two parameters and, unlike most other families, it has a parameter ν that directly controls the differentiability of the process. Larger values of ν correspond to smoother processes. The process is m times mean square differentiable if and only if $\nu > m$ (Stein, 1999). The mean-square differentiability of the process plays a crucial role in kriging prediction. Being able to estimate this parameter from the data gives the model additional flexibility and strength in prediction.

The Matérn family includes the exponential family of correlation functions as a special case with $\nu = 0.5$ and the Gaussian family of correlation function as a limiting case with $\nu \to \infty$. There are several suggested parameterizations of the Matérn family and we follow the one given by Handcock and Wallis (1994):

$$\rho(u; \varphi, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}u}{\varphi} \right)^{\nu} \mathcal{K}_{\nu} \left(\frac{2\nu^{1/2}u}{\varphi} \right),$$

where φ and ν are parameters and $\mathcal{K}_{\nu}$ is the modified Bessel function of order ν as discussed by Abramowitz and Stegun (1965), sec. 9. This parameterization has the nice feature that φ measures how quickly the correlations of the random field decay with distance, and its interpretation is largely independent of ν. Often φ is referred to as the range parameter in the geostatistical literature.

2.2 Kriging, REML, and plug-in method

Kriging is the name for best linear unbiased prediction (BLUP) in the geostatistics literature, see, for example, Cressie (1993) or Stein (1999). When the data can be modeled as in (1) and the parameters of the covariance model $\boldsymbol{\theta}$ are known beforehand, the kriging predictor and kriging

variance at location $\mathbf{s}$ is given by

$$\widehat{Z}(\mathbf{s};\boldsymbol{\theta}) = \boldsymbol{\lambda}^T(\mathbf{s},\boldsymbol{\theta})\mathbf{Z} = \left(\boldsymbol{\Sigma}^{-1}\mathbf{k} + \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x} - \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{k})\right)^T \mathbf{Z}, \tag{2}$$

and

$$M(\mathbf{s};\boldsymbol{\theta}) \;=\; \sigma^2 + \tau^2 - \mathbf{k}^T\boldsymbol{\Sigma}^{-1}\mathbf{k} + (\mathbf{x} - \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{k})^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x} - \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{k}), \tag{3}$$

where $\mathbf{k}$ is the $n \times 1$ vector of covariance between $\mathbf{Z}$ and $Z(\mathbf{s})$, and $\mathbf{x}$ is the $p \times 1$ vector of regressors at location $\mathbf{s}$. The kriging predictor has the minimum mean square prediction error (MSPE) among all linear unbiased predictors.

In practice, however, it is rarely the case that the covariance model is completely known, and often we need to estimate the parameters $\boldsymbol{\theta}$ of the covariance model from the same data from which the predictions are made. When the model completely specifies the likelihood, maximum likelihood (ML) methods can be used to estimate $\boldsymbol{\theta}$. When the ML estimator is in the interior of the parameter space, it is a solution of the score equations. Mardia and Marshall (1984) showed the following result for a consistent sequence of solutions of score equations: When a stationary Gaussian process on $\mathbb{R}^d$ is sampled on an $n_1 \times \ldots \times n_d$ regular lattice with $n_1, \ldots, n_d \to \infty$, and $\rho(\mathbf{u};\boldsymbol{\xi})$ satisfies certain regularity conditions,

$$\mathcal{I}(\boldsymbol{\theta})^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}), \tag{4}$$

where $n = n_1 \times \ldots \times n_d$, $\widehat{\boldsymbol{\theta}}_n$ is the ML estimator of $\boldsymbol{\theta}$ based on $\mathbf{Z}_n = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$, and

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}_\theta\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(\mathbf{Z}_n;\boldsymbol{\theta})\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log p(\mathbf{Z}_n;\boldsymbol{\theta})\right\}^T\right)$$

is the Fisher information matrix for $\boldsymbol{\theta}$ when $\mathbf{Z}_n$ is observed.

Here we use the REML method (Patterson and Thompson, 1971) to estimate $\boldsymbol{\theta}$, which is a ML method based on the likelihood of the contrasts (linear combinations of the data whose mean is zero

for all $\boldsymbol{\beta}$), and has better finite-sample properties (Zimmerman and Zimmerman, 1991). Assuming $\mathbf{X}$ is of full rank, it can be shown that $\mathbf{Y} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}\mathbf{Z}$ forms a basis for all contrasts, and McCullagh and Nelder (1989, p. 247) give the log likelihood for $\boldsymbol{\theta}$ directly in terms of $\mathbf{Y}$:

$$l(\boldsymbol{\theta}) = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log\det\{\boldsymbol{\Sigma}(\boldsymbol{\theta})\} - \frac{1}{2}\log\det\{\mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{X}\} - \frac{1}{2}\mathbf{Y}^T\mathbf{P}(\boldsymbol{\theta})\mathbf{Y}, \tag{5}$$

where $\mathbf{P}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} - \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$. The REML estimator of $\boldsymbol{\theta}$ is obtained by maximizing (5). Cressie and Lahiri (1993) showed that (4) holds for the REML estimator under certain regularity conditions. Under model (1), the (j,k)th element of the Fisher information matrix takes a fairly simple form:

$$\mathcal{I}_{j,k}(\boldsymbol{\theta}) = \frac{1}{2}\mathrm{tr}\left\{\mathbf{P}(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})_j\mathbf{P}(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})_k\right\}, \tag{6}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})_j = \partial\boldsymbol{\Sigma}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_j$.

Let $\widehat{\boldsymbol{\theta}}$ be the REML estimator of $\boldsymbol{\theta}$ based on $\mathbf{Z}$ and $\widehat{Z}(\mathbf{s};\boldsymbol{\theta}) = \boldsymbol{\lambda}^T(\boldsymbol{\theta})\mathbf{Z}$ be the BLUP of $Z(\mathbf{s})$, then $\widehat{Z}(\mathbf{s};\widehat{\boldsymbol{\theta}})$ is the plug-in kriging predictor (also called EBLUP for "empirical" or "estimated" BLUP). We first consider the uncertainty in prediction for EBLUP and give an approximation to its MSPE by taking into account the parameter uncertainty. Since we estimate $\boldsymbol{\theta}$ using only contrasts of $\mathbf{Z}$, $\widehat{Z}(\mathbf{s};\widehat{\boldsymbol{\theta}}) - \widehat{Z}(\mathbf{s};\boldsymbol{\theta})$ is a function of contrasts of $\mathbf{Z}$, which is independent of $\widehat{Z}(\mathbf{s};\boldsymbol{\theta}) - Z(\mathbf{s})$. We have

$$\begin{aligned}
\mathrm{E}(\widehat{Z}(\mathbf{s};\widehat{\boldsymbol{\theta}}) - Z(\mathbf{s}))^2 &= \mathrm{E}(\widehat{Z}(\mathbf{s};\boldsymbol{\theta}) - Z(\mathbf{s}))^2 + \mathrm{E}(\widehat{Z}(\mathbf{s};\widehat{\boldsymbol{\theta}}) - \widehat{Z}(\mathbf{s};\boldsymbol{\theta}))^2 \\
&= M(\mathbf{s};\boldsymbol{\theta}) + \mathrm{E}(\boldsymbol{\lambda}^T(\mathbf{s};\widehat{\boldsymbol{\theta}})\mathbf{Z} - \boldsymbol{\lambda}^T(\mathbf{s};\boldsymbol{\theta})\mathbf{Z})^2,
\end{aligned} \tag{7}$$

where $M(\mathbf{s};\boldsymbol{\theta})$ is the kriging variance given by (3) when the parameters are assumed known (Kackar and Harville, 1984). Harville and Jeske (1992) and Zimmerman and Cressie (1992) suggested that one can approximate the MSPE of the plug-in kriging predictor by

$$V_1(\mathbf{s};\boldsymbol{\theta}) = M(\mathbf{s};\boldsymbol{\theta}) + \mathrm{tr}\left\{\mathcal{I}^{-1}(\boldsymbol{\theta})\left(\frac{\partial\boldsymbol{\lambda}}{\partial\boldsymbol{\theta}}\right)^T\boldsymbol{\Sigma}(\boldsymbol{\theta})\left(\frac{\partial\boldsymbol{\lambda}}{\partial\boldsymbol{\theta}}\right)\right\}, \tag{8}$$

where $\partial\boldsymbol{\lambda}/\partial\boldsymbol{\theta}$ is a matrix with $\partial\lambda_i/\partial\theta_j$ as its i,jth element. Abt (1999) shows by simulation that for exponential correlation functions it is much better than the plug-in estimator of the MSPE in terms of approximating the true MSPE of the EBLUP, and finer approximations considering the correlation between $\widehat{\boldsymbol{\theta}}$ and $\mathbf{Z}$ do not give better results.

2.3 Design Criteria

Our goal is good spatial prediction, including both the point prediction and the prediction interval, over the whole region D. To construct a design criterion, we first construct a pointwise criterion $V(\mathbf{s}, S)$ that measures how well we predict $Z(\mathbf{s})$ conditional on observing the process on S. The design criterion is then defined as a function of $V(\mathbf{s}; S)$ depending on the particular prediction objectives. Two commonly used criteria are:

$$A(S) = \int_D V(\mathbf{s}; S) w(\mathbf{s}) d\mathbf{s} \tag{9}$$

$$M(S) = \sup \{V(\mathbf{s}; S) w(\mathbf{s}) : \mathbf{s} \in D\} \tag{10}$$

where $w(\mathbf{s})$ is the weight assigned to location $\mathbf{s}$. In this paper we only consider the simple case for which $w(\mathbf{s}) \equiv 1$. When the covariance structure is completely known, the MSPE can be easily evaluated by (3) and is usually used as the criterion. In this case, (9) and (10) are referred to as average kriging variance (AKV) and maximum kriging variance (MKV), respectively. However, when the covariance parameters have to be estimated from the data, it is more difficult to define a criterion, as the use of estimated parameters introduces additional uncertainty in both the point prediction and the estimation of the MSPE. The use of estimated parameters often has a larger impact on the estimated MSPE and a lesser effect on the predicted values (Stein, 1999). We seek good point predictions as well as good prediction intervals. Our objective is to find the sampling design that will minimize some combination of the uncertainty in prediction and the uncertainty

in the estimated MSPE, and the critical question is how to incorporate the parameter uncertainty into our design criteria to achieve this objective.

It is obvious that the design criteria will depend on the kind of inference procedure we plan to follow after collecting the data, i.e., the methodology for parameter estimation and prediction at unknown locations. Here we derive design criteria for the "plug-in" method as described in previous section, which is widely used in geostatistics: first estimate the parameter using REML, then "plug in" the estimated parameter into the kriging formula to obtain the plug-in kriging predictor and its MSPE. If we only want good point prediction, V_1 in (8) can be used as the criterion, which is called EK-optimal design in Zimmerman (2005). We will use the notation $V_1(\mathbf{s}; S, \boldsymbol{\theta})$ in the rest of this section to emphasis the dependence of the criterion on the set of sampling locations S.

Next we consider the uncertainty in estimating the MSPE by approximating the variance of the plug-in kriging variance using a second order Taylor expansion of $M(\mathbf{s}; \widehat{\boldsymbol{\theta}})$:

$$\mathrm{Var}\left\{M(\widehat{\boldsymbol{\theta}})\right\} \approx \mathrm{Var}\left\{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \frac{\partial M(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\} \approx \left(\frac{\partial M(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \mathcal{I}^{-1}(\boldsymbol{\theta}) \frac{\partial M(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = V_2(\mathbf{s}; S, \boldsymbol{\theta}). \tag{11}$$

Some linear combination of $V_1(\mathbf{s}; S, \boldsymbol{\theta})$ and $V_2(\mathbf{s}; S, \boldsymbol{\theta})$ can be used as the design criterion for prediction with estimated covariance parameters:

$$V_3(\mathbf{s}; S, \boldsymbol{\theta}, c_1) = V_1(\mathbf{s}; S, \boldsymbol{\theta}) + c_1 V_2(\mathbf{s}; S, \boldsymbol{\theta}), \tag{12}$$

but it is not clear how to choose the weight c_1 appropriately. One can also use (8) as the design criterion under the constraint that (11) is no greater than some chosen constant.

Alteronatively, one can quantify the uncertainty in spatial prediction using functionals of the conditional distribution. Let $p(Z(\mathbf{s})|\mathbf{Z}; \boldsymbol{\theta})$ and $p(Z(\mathbf{s})|\mathbf{Z}; \widehat{\boldsymbol{\theta}})$ be the conditional density of $Z(\mathbf{s})$ at the true and estimated values of the parameter $\boldsymbol{\theta}$, respectively. When $\boldsymbol{\theta}$ is known, $p(Z(\mathbf{s})|\mathbf{Z}; \boldsymbol{\theta})$ is used for inference about Z. The plug-in method uses $p(Z(\mathbf{s})|\mathbf{Z}; \widehat{\boldsymbol{\theta}})$ to make inference when the

parameter is unknown. In order for a design to be good for prediction using the plug-in method, the spread of $p(Z(\mathbf{s})|\mathbf{Z};\boldsymbol{\theta})$ should be small so that the actual conditional variation of the random field is small, and $p(Z(\mathbf{s})|\mathbf{Z};\widehat{\boldsymbol{\theta}})$ should be close to $p(Z(\mathbf{s})|\mathbf{Z};\boldsymbol{\theta})$ so that the uncertainty due to the parameter estimation is small. The conditional variance of $Z(\mathbf{s})$ given $\mathbf{Z}$ evaluated at $\boldsymbol{\theta}$ (i.e., MSPE of the BLUP) is a natural measure for the spread of $p(Z(\mathbf{s})|\mathbf{Z};\boldsymbol{\theta})$. The Kullback divergence of the plug-in conditional density from the conditional density evaluated at $\boldsymbol{\theta}$,

$$D(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}};Z(\mathbf{s})|\mathbf{Z}) = \mathrm{E}\left\{\log\frac{p(Z(\mathbf{s})|\mathbf{Z};\boldsymbol{\theta})}{p(Z(\mathbf{s})|\mathbf{Z};\widehat{\boldsymbol{\theta}})}\right\},$$

can serve as a measure of the distance between the two densities. Define

$$\mathcal{I}(\boldsymbol{\theta};\mathbf{W}|\mathbf{U}) = \mathrm{Cov}_\theta\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(\mathbf{W}|\mathbf{U};\boldsymbol{\theta}), \left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log p(\mathbf{W}|\mathbf{U};\boldsymbol{\theta})\right\}^T\right).$$

We have

$$\mathcal{I}(\boldsymbol{\theta};(\mathbf{W},\mathbf{U})) = \mathcal{I}(\boldsymbol{\theta};\mathbf{U}) + \mathcal{I}(\boldsymbol{\theta};\mathbf{W}|\mathbf{U}),$$

so that $\mathcal{I}(\boldsymbol{\theta};\mathbf{W}|\mathbf{U})$ is the increase in Fisher information for $\boldsymbol{\theta}$ when $\mathbf{W}$ is observed in addition to $\mathbf{U}$. Assuming we use ML or REML to estimate $\boldsymbol{\theta}$, Stein (1999, p. 204) suggested the following two approximations:

$$\begin{aligned} D(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}};Z(\mathbf{s})|\mathbf{Z}) &\approx \frac{1}{2}\mathrm{tr}\left\{\mathcal{I}(\boldsymbol{\theta};\mathbf{Z})^{-1}\mathcal{I}(\boldsymbol{\theta};Z(\mathbf{s})|\mathbf{Z})\right\} &(13)\\ &\approx \frac{\mathrm{E}\{M(\widehat{\boldsymbol{\theta}})-M(\boldsymbol{\theta})\}^2}{4M(\boldsymbol{\theta})^2} + \frac{\mathrm{E}\{\widehat{Z}(\widehat{\boldsymbol{\theta}})-\widehat{Z}(\boldsymbol{\theta})\}^2}{2M(\boldsymbol{\theta})}. &(14)\end{aligned}$$

We can use the first approximation to construct the design criterion

$$V_4(\mathbf{s};S,\boldsymbol{\theta},c_2) = M(\boldsymbol{\theta})\left[1+\frac{c_2}{2}\mathrm{tr}\left\{\mathcal{I}(\boldsymbol{\theta};\mathbf{Z})^{-1}\mathcal{I}(\boldsymbol{\theta};Z(\mathbf{s})|\mathbf{Z})\right\}\right], \quad (15)$$

where c_2 is the turning parameter. Using the second approximation, we can see that only when $c_1 = 1/(2M(\mathbf{s};\boldsymbol{\theta}))$ and $c_2 = 2$, $V_4 \approx V_3$. The weight $c_1 = 1/(2M(\mathbf{s};\boldsymbol{\theta}))$ makes the linear combination

(12) invariant to scale transformations of $\boldsymbol{\theta}$ and $\mathbf{Z}$. We think $c_1 = 1/(2M(\mathbf{s};\boldsymbol{\theta}))$ is a natural choice, and will use it averaged over D as the design criterion in the simulation study:

$$EA(S;\boldsymbol{\theta}) \;=\; \int_D V_3(\mathbf{s};S,\boldsymbol{\theta},1/(2M(\mathbf{s};\boldsymbol{\theta}))d\mathbf{s}. \tag{16}$$

We will refer to it as the estimation adjusted criterion (EA). In general, a large c_1 yields designs that have a heavier emphasis on parameter estimation and give more accurate prediction intervals, while a small c_1 yields designs good for point prediction with estimated parameters. When the range of the covariance function is not too small, these designs are similar to the case in which the parameters are assumed known. Incidentally, Smith and Zhu (2004) reveal a surprising relationship between c_1 and the length of a Bayesian predictive interval with certain coverage probability. For $c_1 = 1/(2M(\mathbf{s};\boldsymbol{\theta}))$, (16) is equivalent to the average length of the Bayesian predictive interval with asymptotic coverage probability of 92%, with larger c_1 corresponding to Bayesian predictive intervals with larger coverage probability. For example, the average length of Bayesian predictive intervals with 90% and 95% coverage probability are equivalent to (16) with $c_1 = 0.41$ and 0.68 respectively. Thus, our choice of c_1 also has a nice Bayesian interpretation. In practice, the choice of the value of c_1 depends on the particular inference one wants to make, and a different value may be preferred depending on the application.

In practice the value of $\boldsymbol{\theta}$ is not known. For fixed $\boldsymbol{\theta}$, we can use $EA(S;\boldsymbol{\theta})$ to find locally optimal designs. When a prior on $\boldsymbol{\theta}$ is available, one can use the following Bayesian design criterion:

$$EA^B(S) = \int_\Theta EA(S;\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Alternatively, one can define the relative efficiency for EA criterion as

$$RE(S;\boldsymbol{\theta}) = \frac{EA(S;\boldsymbol{\theta})}{EA(S(\boldsymbol{\theta});\boldsymbol{\theta})}, \tag{17}$$

which measures the relative performance of design S with respect to the locally optimal design $S(\boldsymbol{\theta}) = \arg\min EA(S;\boldsymbol{\theta})$ at $\boldsymbol{\theta}$, and find the minimax design S_M such that

$$S_M = \arg\min_{S\subset D} \max_{\boldsymbol{\theta}\in\Theta} RE(S;\boldsymbol{\theta}). \tag{18}$$

Both approaches are rather computationally intensive. A simple example of minimax design is given in Section 4.

3 Numerical Algorithms

In this section we first describe the two-step algorithm we developed to find the best spatial sampling design for the EA criterion, followed by a discussion of the simulation method we used in our simulation studies. At the end of this section we outline how we compare the effectiveness of different designs.

3.1 Two-step Algorithm

In the spatial sampling design problems we consider, we need to choose n sample points from a fine grid of allowable sampling locations D of size N that minimize some objective function. Since the objective functions we consider have complicated functional forms and involve integrals that cannot be explicitly evaluated, they are fairly expensive to compute. Furthermore, N is usually large enough so that there is no exact algorithm available at present that could find the optimal solution even if the objective function were easy to calculate.

For small sample sizes we can use heuristic algorithms such as SA (Zhu and Stein, 2005) to find the optimal design. When the sample size is moderately large, the SA fails to deliver reasonable results, and we propose a "two-step" algorithm for those occasions. Observing that for the problem of finding designs for prediction with estimated parameters, the design criteria we propose are

effectively some compromise between prediction and parameter estimation, we suggest using the following algorithm to find an approximate solution of the original optimization problem:

- Step one: For a fixed proportion $p \in (0,1)$, find an optimal size $(1-p)n$ design for prediction with given parameters, and then choose pn additional points using SA so that the combined sample of size n optimizes an estimation-based criterion.

- Step two: For different values of p, compute the EA values for the combined samples of size n thus obtained, and find the one that minimizes the EA criterion.

The computational advantages this algorithm has over direct minimization of the original objective function are: first, the design criteria for prediction or estimation alone are much easier to compute than the design criteria for prediction with estimated parameters. Second, when $(1-p)n$ is large, based on theoretical and simulation studies, we know that a regular design is good for prediction. So in practice it is usually enough to find the optimal design for prediction among a sequence of regular designs, which can dramatically reduce the computational time. Last, our limited experience indicates that the designs that minimize the objective function usually have $p \ll 1/2$ for n moderately large, thus we only need to consider size pn designs for estimation with $pn \ll n$, which usually can be handled by SA rather efficiently. In the simulation studies we use the AKV of the prediction as the design criterion for prediction and the logarithm of the determinant of the inverse of Fisher information matrix (LDF) as the design criterion for estimation.

3.2 Simulation Method

We use the following setting in our simulation studies: The region of interest is taken to be the unit square $[0,1]^2$. The allowable sampling grid D in most cases is $\{0, 0.01, \ldots, 1\}^2$. The evaluation grid E, on which the objective function is evaluated, is $\{0.005, 0.015, \ldots, 0.995\}^2$, which is the center of

each of the smallest squares in grid D. In some cases when the sample size is not large, a subset of E is used instead to speed up the computation.

Two algorithms are used to find the optimal designs. For small sample size we use SA. For moderately large to large sample size designs for prediction with unknown parameter vector, we use the two-step algorithm. For each design the objective function is evaluated at the evaluation grid E and the integral in (9) is approximated by summation.

We use simulations to compare designs obtained by different criteria to demonstrate the effectiveness of the criterion. The random samples are simulated using model (1) with mean zero and Matérn family of correlation function with preselected parameters. The Choleski decomposition method is used to simulate the data with correct covariance structure as specified by the model. Only those process values on the design sites are simulated. We then use the REML method to estimate the parameters from the simulated data and use the plug-in method to obtain the kriging predictors $\widehat{Z}(\mathbf{s};\widehat{\boldsymbol{\theta}})$ and kriging variances $M(\mathbf{s};\widehat{\boldsymbol{\theta}})$ at the evaluation grid E using the estimated parameters. The numerical algorithm for obtaining the REML estimates is based on the method described in Mardia and Marshall (1984) with some modification. See Zhu and Stein (2005) for more details.

3.3 Design Evaluation

To compare the effectiveness of designs, we need to consider both the accuracy of the predictions (usually measured by the true mean square error) on the evaluation grid E and the accuracy of the estimated MSPEs of the predictions. The distribution of the random process on the evaluation grid E, conditional on the observed data at the design sites S, is normal with conditional mean and variance given by, respectively, the kriging predictor $\widehat{Z}(\mathbf{s};\boldsymbol{\theta})$ and kriging variance $M(\mathbf{s};\boldsymbol{\theta})$

evaluated at the true parameter values. The true conditional mean square prediction error for the ith simulation at site $\mathbf{s}$ is $\text{MSPE}^{(i)}(\mathbf{s}) = (\widehat{Z}^{(i)}(\mathbf{s};\widehat{\boldsymbol{\theta}}^{(i)}) - \widehat{Z}^{(i)}(\mathbf{s};\boldsymbol{\theta}))^2 + M(\mathbf{s};\boldsymbol{\theta})$, where $\widehat{Z}^{(i)}$ and $\widehat{\boldsymbol{\theta}}^{(i)}$ are the kriging predictor and the REML estimator of $\boldsymbol{\theta}$ for the ith simulation respectively. Thus we can average $\text{MSPE}^{(i)}(\mathbf{s})$ across simulations to get the true MSPE of the plug-in kriging prediction. Note that there is no need to simulate $Z(\mathbf{s})$ at locations outside the design sites to evaluate the conditional MSPE. Some weighted average of the true MSPE thus obtained over E can be used as a measure of the accuracy of prediction for a certain design. Simulation from an estimated model can also be used to obtain a more accurate assessment of MSPE in a prediction problem and is usually referred to as parametric bootstrapping (Davison and Hinkley, 1997).

Measuring the accuracy of the estimated MSPE deserves more consideration. The measures we considered for the accuracy of MSPE estimator $M(\mathbf{s};\widehat{\boldsymbol{\theta}})$ are:

1. MSE of $M(\mathbf{s};\widehat{\boldsymbol{\theta}})$: $\text{E}(M(\mathbf{s};\widehat{\boldsymbol{\theta}}) - M(\mathbf{s};\boldsymbol{\theta}))^2$
2. MSE of Ratio: $\text{E}(M(\mathbf{s};\widehat{\boldsymbol{\theta}})/M(\mathbf{s};\boldsymbol{\theta}) - 1)^2$
3. Mean Square Log Ratio (MSLR): $\text{E}\left(\log(M(\mathbf{s};\widehat{\boldsymbol{\theta}})/M(\mathbf{s};\boldsymbol{\theta}))\right)^2$
4. Gamma Deviance: $\text{E}\left(-\log(M(\mathbf{s};\widehat{\boldsymbol{\theta}})/M(\mathbf{s};\boldsymbol{\theta})) + (M(\mathbf{s};\widehat{\boldsymbol{\theta}}) - M(\mathbf{s};\boldsymbol{\theta}))/M(\mathbf{s};\boldsymbol{\theta})\right)$

The MSE of the MSPE is a commonly used measure, which measures the average squared distance of the estimated MSPE to the true MSPE. However, we believe it penalizes overestimates of MSPE more heavily than it does underestimates of MSPE. For example, if the variance of $M(\mathbf{s};\widehat{\boldsymbol{\theta}})$ compared to the magnitude of the MSPE is small, then the maximum penalty for underestimating the MSPE is about $M(\mathbf{s};\boldsymbol{\theta})^2$, while for overestimating the MSPE the penalty is unbounded.

The MSE of the ratio $M(\mathbf{s};\widehat{\boldsymbol{\theta}})/M(\mathbf{s};\boldsymbol{\theta})$ does not depend on the magnitude of the MSPE of prediction, and the value can be interpreted as the percentage difference between $M(\mathbf{s};\widehat{\boldsymbol{\theta}})$ and MSPE. However, it also penalizes overestimation more severely than underestimation.

The MSLR does not depend on the magnitude of the MSPE of prediction, and it penalizes overestimation and underestimation equally on the log scale. We have observed empirically that the MSLR has approximately a Gaussian distribution. However, its interpretation is less transparent.

The gamma deviance is used because of the empirical observation that $M(\mathbf{s}; \widehat{\boldsymbol{\theta}})$ has approximately a gamma distribution.

Among these criteria, (2) and (4) are locally equivalent (i.e., when $M(\mathbf{s}; \widehat{\boldsymbol{\theta}})$ is close to MSPE, the difference between (2) and (4) is of higher order in terms of $1 - M(\mathbf{s}; \widehat{\boldsymbol{\theta}})/M(\mathbf{s}; \boldsymbol{\theta})$). There is no compelling reason to prefer any one of the above measures, and in the simulation studies we conducted all measures yield similar results. We will present the results using MSE of the ratio as the measure of accuracy of $M(\mathbf{s}; \widehat{\boldsymbol{\theta}})$.

4 Simulation Results

In this section we present the results of our simulation studies. In the first simulation study, SA is used to find the best spatial designs, and our results show that designs using EA criterion (D_{EA}) give much better estimators of MSPE than those using AKV criterion (D_{AKV}), while the true MSPE for the two types of designs are similar. A minimax version of D_{EA} is also compared with D_{AKV}, and the results are similar. In the second simulation study, we compare the performance of the two-step algorithm with SA for finding best D_{EA} with sample size $n = 100$. The two-step algorithm finds better designs in shorter time.

In all the simulation studies we assumed that the random processes have mean zero, and focused on the more interesting problem of accounting for uncertainty in the estimation of the covariance parameters $\boldsymbol{\theta}$ on the prediction of $\mathbf{Z}(\mathbf{s})$.

4.1 Simulation for Prediction with Unknown Parameters: Small Sample Size

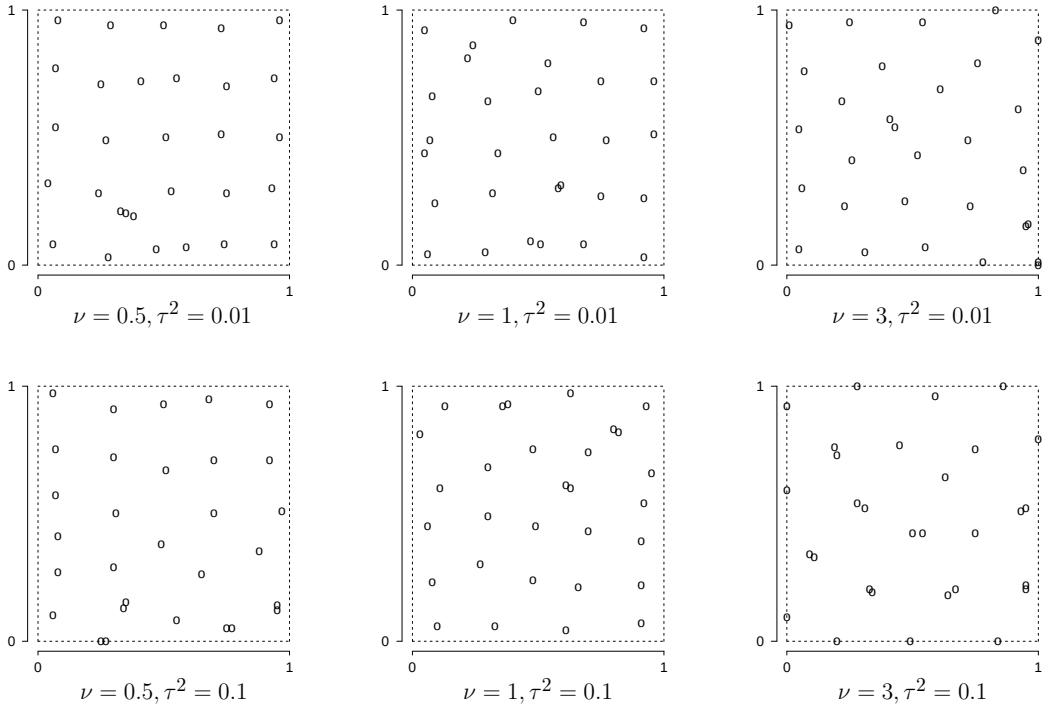

Figure 1: Plots of designs for prediction with unknown parameters. For all the plots $\varphi = 0.5$ and $\sigma^2 = 1$.

In this simulation study the parameters are assumed unknown and have to be estimated from the data. The region of interest is the unit square $[0, 1]^2$ and the objective is to predict the random field on the evaluation grid E. We considered cases where the sample size $n = 30$, and the true parameter values were $\varphi = 0.5$, $\nu \in \{0.5, 1, 3\}$, $\sigma^2 = 1$ and $\tau^2 \in \{0.01, 0.1\}$. We used the EA design criterion (16) evaluated at the true parameter value. SA was used to search for the optimal sampling design. The resulting designs are plotted in Figure 1, which shows that the designs exhibit some clustering in addition to a rather regularly spaced design, and the amount of clustering is dependent on the true parameter values.

Table 1: Comparison of D_{EA} (estimation adjusted, (16)) and D_{AKV} (average kriging variance) designs using simulation

(ν, τ^2)	$(0.5, 0.01)$		$(1.0, 0.01)$		$(1.0, 0.10)$		$(3.0, 0.10)$	
Design	D_{EA}	D_{AKV}	D_{EA}	D_{AKV}	D_{EA}	D_{AKV}	D_{EA}	D_{AKV}
True MSPE	0.111	0.108	0.109	0.115	0.260	0.253	0.167	0.167
MSE of Ratio	0.273	1.273	0.438	1.318	0.217	0.317	0.159	0.270

Notes: Results are based on 500 simulations. For all designs $\varphi = 0.5$, $\sigma^2 = 1$ and the sample size $n = 30$.

Table 2: Comparison of Minimax and D_{AKV} designs using simulation

	$\nu = 1.0, \tau^2 = 0.01$		$\nu = 1.0, \tau^2 = 0.10$		$\nu = 3.0, \tau^2 = 0.10$	
Design	Minimax	D_{AKV}	Minimax	D_{AKV}	Minimax	D_{AKV}
True MSPE	0.112	0.117	0.258	0.254	0.167	0.167
MSE of Ratio	0.365	1.299	0.275	0.362	0.222	0.262

Notes: Results are based on 500 simulations. For all designs $\varphi = 0.5$, $\sigma^2 = 1$ and the sample size $n = 30$.

To compare the practical performance of the designs, we simulated 500 Gaussian random fields with the Matérn correlation function for each parameter combination in the study using the locations shown in Figure 1 as well as in a D_{AKV} design (not shown here). For each simulation, the parameters were estimated using REML using observations in either the D_{EA} design or the D_{AKV} design. We then use the plug-in procedure to obtain the plug-in kriging predictor and the plug-in kriging variance at the evaluation grid E. The true MSPE of the plug-in kriging prediction and the MSE of the ratio $M(\mathbf{s}; \widehat{\boldsymbol{\theta}})/M(\mathbf{s}; \boldsymbol{\theta})$ are shown in Table 1. The table shows that the true MSPE of the plug-in prediction for D_{EA} and D_{AKV} designs are very close, but the D_{EA} designs give better estimates of MSPE. The improvement is always substantial and sometimes dramatic.

In practice, the true parameters of the process are unknown and it is desirable to have a design that has good performance for a range of parameters. As a simple example, we calculated the relative efficiency for the six parameter combinations considered in Figure 1, and found that the design for $\nu = 0.5$ and $\tau^2 = 0.01$ minimizes the maximum relative efficiency among the six designs, which we will refer to as the minimax design. The maximum relative efficiency is 1.045, meaning

Table 3: Comparison of Two-step algorithm and SA using different design criteria

Design:		Two-step							SA
	p	100.0 %	20.0%	13.3%	10.0%	6.7%	3.3%	0.0%	NA
$n = 30$	EA	304.9	127.3	126.8	125.2	122.0	121.6	3908.2	118.0
	AKV	273.7	110.0	104.3	102.3	99.3	96.3	93.2	101.1
	LDF	-17.6	-16.0	-15.7	-15.4	-15.1	-14.4	-7.3	-14.8
	p	100.0%	10.0%	7.0%	5.0%	3.0%	1.0%	0.0%	NA
$n = 100$	EA	95.5	42.9	42.3	41.9	42.2	45.0	84.6	42.5
	AKV	93.1	41.3	40.5	39.9	39.5	39.1	38.9	40.2
	LDF	-21.9	-20.4	-20.1	-19.8	-19.4	-18.4	-16.2	-19.6

Notes: EA refers to the EA criterion (16); AKV represents Average Prediction Variance assuming the parameters are known; LDF represents Logarithm of the Determinant of the Fisher information matrix. The value of both EA and AKV are multiplied by 1000. All of them are evaluated at the true parameter values. p is the percentage of points assigned for parameter estimation in D_{TS} designs.

that, in the worst case, the EA criterion for the minimax design is 4.5% larger than that for the best design we obtained for the true parameter values. In Table 2, simulation results for the minimax design are compared with the D_{AKV} design for several different parameters. In all cases the minimax design gives substantially better estimates of the MSPE while the MSPE of prediction is similar to that of the D_{AKV} design.

4.2 Simulation for the Two-step Algorithm

In this section we study the two-step algorithm as described in Section 3.1 and compare them with SA. The D_{AKV} and D_{LDF} (designs for parameter estimation using LDF as the criterion) designs are two extreme cases in the two-step algorithm with $p = 0$ and 1, respectively, which are also included in the comparison. Sample sizes $n = 30$ and $n = 100$ were considered in this study. For both $n = 30$ and $n = 100$, ten other designs were considered besides the D_{AKV} and D_{LDF} designs in the second step, with the number of points assigned to parameter estimation ranging from one to ten, respectively. For each of these designs, we first find a size $(1-p)n$ design for prediction with given parameters using SA and a regular design as starting point. Then we fix these $(1-p)n$ points and use SA to add pn points that minimize the LDF of the combined size n design. We ran SA

50 times using different random starting points and chose the best one from them. Some of these designs are compared with designs found by SA in Table 3. In all the designs the true parameter values were assumed to be $\varphi = 0.5$, $\nu = 1$, $\sigma^2 = 1$ and $\tau^2 = 0.01$. From the table we can see that for $n = 30$, assigning one point for estimation ($p = 3.3\%$) gives the best D_{EA} design, which is still worse than the design found by SA. Since the two-step algorithm searches over a smaller design space, when the sample size is small and SA can find designs very close to optimal, we expect the design found by the two-step algorithm to be worse. Nevertheless the difference between these two is not big.

For $n = 100$, assigning 5 points for estimation ($p = 5\%$) minimizes the EA criterion, which maintains a balance between prediction and parameter estimation. From Table 3 we can see that this design is even slightly better than the design found by SA. This is because for even moderately large sample size, the optimization problem becomes so complex that SA with multiple random starting designs may not be able to find designs that are very close to optimal. The two-step algorithm, which exploited the special structure of this problem, can find good designs by searching a much smaller design space. Furthermore, the time it takes is only a fraction of that of SA. In our simulation study, it takes 245,069 seconds (about 68 hours) to run the SA once on a Pentium IV 1.7G PC, while it only takes 55,061 seconds (about 15 hours) to run the two-step algorithm with the second step using multiple starting designs. Figure 2 gives two D_{EA} designs found by the two-step algorithm for sample size $n = 30$ and $n = 100$ respectively. From it we can see that the second step adds points that are very close to each other, which is a typical feature in designs for parameter estimation and is helpful for estimating both the nugget effects and the smoothness parameters. Note that the designs found by both SA and the two-step algorithm are dependent on the initial design and are by no means unique. For different initial designs, both algorithms

can give designs with quite different spatial configurations, while in terms of criterion value the two-step algorithm is less sensitive to the initial designs.

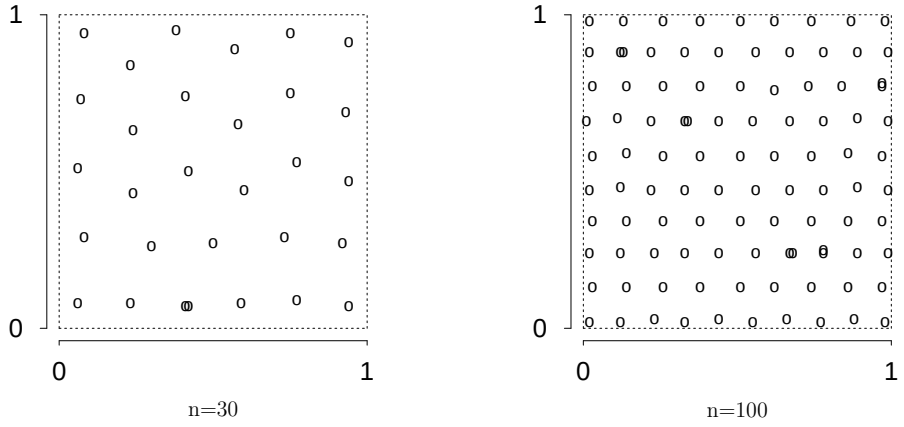

Figure 2: Plot of D_{EA} designs found by the two-step algorithm with sample size $n = 30$ and $n = 100$ respectively.

For $n = 100$, six of the designs in Table 3 were compared using simulation in a similar manner as in Table 1; the results are in Table 4. The D_{EA} found by the two-step algorithm ($p = 5.0\%$) has the smallest average MSPE and estimates MSPE better than the D_{EA} found by SA or D_{AKV} ($p = 0\%$). This is consistent with the ordering of the EA criterion value given in Table 3, giving some evidence that it is a reasonable design criterion to use. It also confirms the effectiveness of the two-step algorithm as an alternative to SA. Compared to D_{AKV} designs, not only can assigning a small proportion of points for parameter estimation substantially improve the MSPE estimators, but it can even reduce the prediction error of the plug-in kriging predictors.

MSE of the ratio, the measure of the accuracy of MSPE, has the same relative ordering as the relative ordering of the LDF value for each design (see Table 3). Since the relative ordering of LDF has been shown to represent the variance of the estimated parameters, these results offer evidence

Table 4: Comparison of designs found by two-step algorithm and SA using simulation

Design:	Two-step					SA
p	100.0%	10.0%	5.0%	1.0%	0.0%	NA
MSPE of Prediction	0.0941	0.0425	0.0407	0.0413	0.0417	0.0408
MSE of Ratio	0.0500	0.0693	0.0858	1.0122	1.7711	0.4769

Notes: Results are based on 100 simulations. Sample size $n = 100$. p is the percentage of points assigned for parameter estimation in D_{TS} designs.

that parameter estimation is closely related to the estimation of MSPE.

5 Example: Redesigning an Air Monitoring Network

In this section, we consider the problem of optimally reducing a SO_2 monitoring network to maintain accurate spatial prediction. The monitoring network we consider covers the geographic area of IL, IN, OH, and KY, with 101 stations monitoring SO_2 in this area in 2002. The locations of these stations are shown on the left of Figure 3. We sought the best way to reduce the network to 50 stations which gives the best spatial prediction in terms of both MSPE and the variance of the estimated MSPE. The data we used to fit the spatial model are the 2002 annual mean of the SO_2 level in the unit of parts per billion, downloaded from http://www.epa.gov/air/data/index.html. Baumgardner et al. (1999) and Baumgardner et al. (2002) provide descriptions of the measurement methods used and some summary statistics. Recent research on SO_2 data have focused on the estimation of the temporal trend (Malm et al., 2002; Mueller, 2003; Holland et al., 2004), and very few attempts have been made to fit spatial models for SO_2 concentration data. Holland et al. (2004) use a spatial model for estimating the regional trend of SO_2, in which their response variable is the estimated site-specific trend rather than the concentration. Here we are interested in the spatial prediction of the annual mean of SO_2 concentration. Preliminary data analysis shows that there is a high concentration along the Ohio river valley, and a logarithmic transformation makes the distribution of the data closer to normal. We first fit model (1) to the logarithm of 2002 annual

mean of the SO_2 level in the unit of parts per billion, with the mean assumed to be a polynomial function of the longitude (s_1) and latitude (s_2) of the locations of the stations, and covariance function from the Matérn family of covariance functions with nugget. Only s_1, s_2, and s_2^2 turn out to be significant in the mean model, and the variogram of the residuals shows no evidence of nonstationarity or anisotropy. The REML estimator of ν tends to infinity, indicating that the Gaussian covariance function $C(u;\boldsymbol{\theta}) = \sigma^2 \exp(-u^2/\varphi^2) + \tau^2$ is appropriate here. As a result we use the following model:

$$\log \mathrm{SO}_2(\mathbf{s}) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 s_2^2 + \epsilon(\mathbf{s}),$$

where $\epsilon(\mathbf{s})$ is a stationary Gaussian random field with isotropic Gaussian covariance function. Using restricted maximum likelihood we obtained the estimates $\widehat{\varphi} = 0.28$, $\widehat{\sigma}^2 = 0.093$ and $\widehat{\tau}^2 = 0.045$. These estimated parameter values are used to evaluate the design criteria, so the design we have here is a locally optimal design, assuming that the covariance parameters will not change dramatically from year to year. Only the most recent year's data were used to estimate the parameters, because we do not want to assume the covariance structure is fixed in time. It is possible to fit multiple years of data with a spatial-temporal model allowing a dynamic change of the covariance parameters and use it to predict the covariance parameters of the current year. Though this method has the potential of offering better estimates of the covariance parameters for evaluating the design criteria, we did not pursue it because of the purely spatial emphasis of this paper. It is also conceivable that a spatial non-stationary model might give a better fit to this data. Given that the region we considered only covers four states with similar geographical features, though, it is unlikely to make a significant difference, and we do not pursue this possibility here. It is worth pointing out that if the design were for a larger geographical area, the use of non-stationary models would need to be considered.

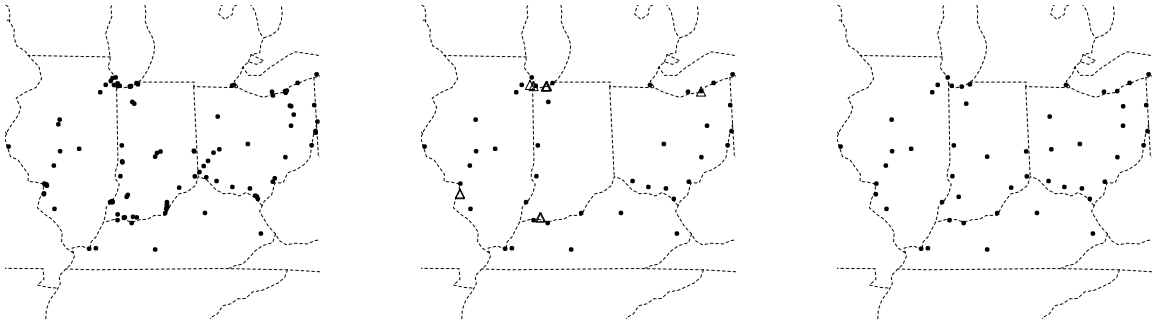

Figure 3: Plot of air monitoring network in EPA Region 5 (Left), D_{TS} design (middle), and D_{AKV} design (right). D_{TS} and D_{AKV} designs are of size $n = 50$. In the D_{TS} design, the triangles are the 10 sites chosen to optimize the LDF criterion.

Using the two-step algorithm described in Section 3.1, we found the D_{TS} design plotted in the middle of Figure 3, in which 40 sites are used to optimize the prediction-based criterion AKV (black dots), and ten sites are used to optimize the estimation-based criterion LDF (triangles), which gives the smallest value for the EA criterion. Among the ten sites for estimation, there are two sites very close to each other in both south IL and south IN, and another three very close sites at the bottom of Lake Michigan. For comparison purposes, we also find the D_{AKV} design that minimizes AKV, assuming that the covariance parameters are known, which is plotted on the right in Figure 3. Note that in the D_{EA} design, the ten sites selected for estimation are either very close to each other or close to an existing site; these sites yield important information for estimating the covariance parameters, while the D_{AKV} design minimizing AKV is rather similar to a space-filling design. There are relatively more points near the boundary in both designs though, because of the mean structure we have in the model.

Table 5 gives the comparison between the D_{EA} design and the D_{AKV} design we found for

Table 5: Comparison of D_{TS} and D_{AKV} design of air monitoring network

Design	MSPE of Prediction	MSE of Ratio
D_{EA}	0.147	0.061
D_{AKV}	0.149	0.103

Notes: Results are based on 100 simulations. For all designs $\widehat{\varphi} = 0.28$, $\widehat{\sigma}^2 = 0.093$ and $\widehat{\tau}^2 = 0.045$.

reducing the air monitoring network. Although both designs give point predictors with similar accuracy, with the MSPE for D_{EA} design slightly better, it is in the area of estimating MSPE that the D_{EA} design really makes a difference, with the MSE of ratio much smaller than that of the D_{AKV} design.

6 Discussion

We propose using a new EA design criterion for prediction with estimated parameters and the two-step algorithm to find a design that optimizes this criterion approximately when the sample size is moderately large. Our simulation studies show that when we have to use the same data for both estimation and prediction, the D_{EA} designs that minimize the EA criterion assign a small proportion of points ($3 - 10\%$ in our examples) for estimation, and it is better than the regular square/triangular designs that are optimal for prediction with known parameters. The improvement is very limited in terms of the prediction error, but can be large in terms of estimating MSPE.

The plug-in method is considered in this paper as the inference procedure because it is a very common method in practice, and the plug-in estimator of the kriging variance is used to estimate the MSPE. Zimmerman and Cressie (1992) propose to use a different estimator for estimating the prediction error when the spatial correlation is known to be weak, and Diggle and Ribeiro (2002) among others promote the use of Bayesian inference. It would be interesting to study what effect different inference procedures may have on the sampling design and whether our design criterion is

good for inference procedures other than the plug-in method.

In particular, Bayesian prediction provides a natural way to account for the effect of parameter uncertainty on prediction uncertainty by averaging over the posterior distribution of parameters to get the predictive distribution. It would be interesting to derive design criteria based on functionals of the Bayesian predictive distribution. It is, however, computationally infeasible to carry out a brute force Bayesian calculation in this context, as there is no closed form solution for the predictive distribution and some MCMC method or numerical integration has to be used. We need to generate the predictive distributions for a large number of sites in the region D to adequately evaluate the performance of different designs. Furthermore, generally many designs have to be evaluated before a good design can be found. Since one needs to be able to examine many designs quickly, it is not feasible to use a long MCMC to evaluate a single design. Some type of asymptotic approximation to the predictive distribution is needed to make progress.

In our simulation studies, the design criteria are always evaluated at the true parameter values used for simulation, as our design criteria are dependent on the parameter values. A minimax criterion can be used on a discretized parameter space instead. If one can appropriately put a prior on the parameter space, one can use our criterion averaged over the prior distribution as the final design criterion. We have to be very cautious about how to choose the prior though, as it will have a huge influence on the design and there is no data available to lessen its effect. Both approaches are more computationally intensive than what we have done here.

In this paper we discuss the spatial sampling design problem under the assumption that the data can be modeled as observations from a Gaussian random field with stationary covariance structure. Both Gaussian and stationarity assumptions may not be satisfied in practice. It is of interest to investigate designs for non-Gaussian random fields and for non-stationary covariance

structure. For non-Gaussian random fields, one can make inference using the spatial generalized linear mixed model (Diggle et al., 1998). Unlike in the Gaussian case, there is no closed form formula to calculate the MSE even when the covariance parameters can be assumed known, and some approximation is needed to derive appropriate design criteria parallel to the Gaussian case. We intend to study this in a seperate paper. When the covariance structure is non-stationary, the design problem becomes more complex, as the estimation of the non-stationarity structure adds yet another source of uncertainty. Müller-Gronbach and Ritter (1998) show that nonadaptive designs are less efficient than adaptive designs for prediction of one dimensional Gaussian random processes with inhomogeneous local smoothness. Unless the nature of the nonstationarity is known ahead of time, some adaptive design scheme may be necessary for such problems.

References

Abramowitz, M. and Stegun, I. (1965), *Handbook of Mathematical Functions, ninth ed.*, New York: Dover.

Abt, M. (1999), "Estimating the prediction mean squared error in Gaussian stochastic processes with exponential correlation structure," *Scandinavian Journal of Statistics*, 26, 563–578.

Banjevic, M. and Switzer, P. (2002), "Bayesian Network Designs for fields with variance as a function of the location," *Proceedings of the 2002 JSM Conference*.

Baumgardner, R., Isil, S., Bowser, J., and Fitzgerald, K. (1999), "Measurements of rural sulfur dioxide and particle sulfate: Analysis of CASTNet data, 1987 through 1996," *Journal of the Air & Waste Management Association*.

Baumgardner, R., Lavery, T., Rogers, C., and et al. (2002), "Estimates of the atmospheric de-

position of sulfur and nitrogen species: Clean Air Status and Trends Network, 1990-2000," *Environmental Science & Technology*.

Benhenni, K. and Cambanis, S. (1992), "Sampling designs for estimating integrals of stochastic processes," *The Annals of Statistics*, 20, 161–194.

Caselton, W. and Zidek, J. (1984), "Optimal monitoring network designs," *Statistics and Probability letters*, 2, 223–227.

Caselton, W. F., Kan, L., and Zidek, J. V. (1992), "Quality data networks that minimize entropy," in *Statistics in the Environmental and Earth Sciences*, pp. 10–38.

Cressie, N. (1993), *Statistics for Spatial Data: revised ed*, New York: John Wiley.

Cressie, N. and Lahiri, S. N. (1993), "The asymptotic distribution of REML estimators," *Journal of Multivariate Analysis*, 45, 217–233.

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap methods and their application*, New York: Cambridge University Press.

Diggle, . P. J. and Ribeiro, P. J. (2002), "Bayesian inference in Gaussian model based geostatistics," *Geographical and environmental modelling*.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based geostatistics (Disc: p326-350)," *Applied Statistics*, 47, 299–326.

Fisher, R. (1925), "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*.

Guttorp, P., Le, N. D., Sampson, P. D., and Zidek, J. V. (1993), "Using entropy in the redesign of an environmental monitoring network," in *Multivariate Environmental Statistics*, pp. 175–202.

Handcock, M. S. and Wallis, J. R. (1994), "An approach to statistical spatical-temporal modeling of meteorological fields," 89, 368–378.

Harville, D. A. and Jeske, D. R. (1992), "Mean squared error of estimation or prediction under a general linear model," *Journal of the American Statistical Association*, 87, 724–731.

Holland, D., Caragea, P., and Smith, R. (2004), "Regional trends in rural sulfur concentrations," *Atmospheric Environment*.

Kackar, R. N. and Harville, D. A. (1984), "Approximations for standard errors of estimators of fixed and random effects in mixed linear models," *Journal of the American Statistical Association*, 79, 853–862.

Ko, C., Lee, J., and Queyranne, M. (1995), "An exact algorithm for maximum entropy sampling," *Operations Research*, 43, 684–691.

Lark, R. (2002), "Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood," *Geoderma*, 105, 49–80.

Lee, J. (1998), "Constrained maximum-entropy sampling," *Operations Research*, 46, 655–664.

Malm, W., Schichtel, B., Ames, R., and et al. (2002), "A 10-year spatial and temporal trend of sulfate across the United States," *Journal of Geophysical Research-Atmospheres*.

Mardia, K. V. and Marshall, R. J. (1984), "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika*, 71, 135–146.

McBratney, A. B., Webster, R., and Burgess, T. M. (1981), "The design of optimal sampling schemes for local estimation and mapping of regionalized variables, I – Theory and method," *Computers and Geosciences*, 7, 331–334.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models (Second edition)*, Chapman & Hall Ltd.

Mueller, S. (2003), "Seasonal aerosol sulfate trends for selected regions of the United States," *Journal of the Air & Waste Management Association*.

Müller-Gronbach, T. and Ritter, K. (1998), "Spatial adaption for predicting random functions," *The Annals of Statistics*, 26, 2264–2288.

Patterson, H. D. and Thompson, R. (1971), "Recovery of inter-block information when block sizes are unequal," *Biometrika*, 58, 545–554.

Ritter, K. (1996), "Asymptotic optimality of regular sequence designs," *The Annals of Statistics*, 24, 2081–2096.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and analysis of computer experiments (C/R: p423-435)," *Statistical Science*, 4, 409–423.

Smith, R. L. and Zhu, Z. (2004), "Asymptotic theory for kriging with estimated parameters and its application to network design," Tech. rep., UNC-Chapel Hill.

Stein, M. L. (1995), "Locally Lattice Sampling Designs for Isotropic Random Fields," *The Annals of Statistics*, 23, 1991–2012.

— (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer-Verlag.

van Groenigen, J. W. and Stein, A. (1998), "Constrained optimization of spatial sampling using continuous simulated annealing," *Journal of Environomental Quality*, 43, 684–691.

Wiens, D. P. (2004), "Robustness in Spatial Studies I: Minimax Design," *Environmetrics*, In press.

Wikle, C. K. and Royle, J. A. (1999), "Space-time dynamic design of environmental monitoring networks," *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 489–507.

Yfantis, E. A., Flatman, G. T., and Behar, J. V. (1987), "Efficiency of kriging estimation for square, triangular, and hexagonal grids," *Mathematical Geology*, 19, 183–205.

Zhu, Z. and Stein, M. (2005), "Spatial Sampling Design for Parameter Estimation of the Covariance Function," *Journal of Statistical Planning and Inference*, In press.

Zidek, J., Sun, W., and Le, N. (2000), "Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields," 49, 63–79.

Zimmerman, D. L. (2005), "Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction," *Environmetrics*.

Zimmerman, D. L. and Cressie, N. (1992), "Mean squared prediction error in the spatial linear model with estimated covariance parameters," *Annals of the Institute of Statistical Mathematics*, 44, 27–43.

Zimmerman, D. L. and Zimmerman, M. B. (1991), "A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors," *Technometrics*, 33, 77–91.